Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W567–W569***
doi:10.1093/nar/gki388

FeatureExtract—extraction of sequence annotation made easy

Rasmus Wernersson*

Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark

Received February 14, 2005; Revised and Accepted March 10, 2005

ABSTRACT

Work on a large number of biological problems benefits tremendously from having an easy way to access the annotation of DNA sequence features, such as intron/exon structure, the contents of promoter regions and the location of other genes in upsteam and downstream regions. For example, taking the placement of introns within a gene into account can help in a phylogenetic analysis of homologous genes. Designing experiments for investigating UTR regions using PCR or DNA microarrays require knowledge of known elements in UTR regions and the positions and strandness of other genes nearby on the chromosome. A wealth of such information is already known and documented in databases such as GenBank and the NCBI Human Genome builds. However, it usually requires significant bioinformatics skills and intimate knowledge of the data format to access this information. Presented here is a highly flexible and easy-to-use tool for extracting feature annotation from GenBank entries. The tool is also useful for extracting datasets corresponding to a particular feature (e.g. promoters). Most importantly, the output data format is highly consistent, easy to handle for the user and easy to parse computationally. The FeatureExtract web server is freely available for both academic and commercial use at http://www.cbs.dtu.dk/services/FeatureExtract/.

INTRODUCING THE ANNOTATION STRING

Central to the way FeatureExtract processes annotation information is a simple but powerful concept—the use of an 'annotation string', which is a string of characters the same length as the DNA string. Each position in the annotation string describes the corresponding position in the DNA string. This is best illustrated by a short example showing how the start of the yeast gene YBR062C will be annotated:

```
Sequence:   ATGTCTACATATGAAGGTATGTAA
Annotation: (EEEEEEEEEEEEEE)DIIIIIII
```

The first 16 bp are the first exon—annotated with 'E's and parentheses to indicate the beginning and end of the exon block. The seventeenth position is the very first position in the first intron—the donor site, annotated with the letter 'D'. The rest of the intron is annotated with 'I's until the last position (the acceptor site) is reached, which is annotated with an 'A' (not shown here).

The FeatureExtract tool has built-in support for working with not only protein-coding genes (exon/intron annotation), but also a wide range of other types of sequences, including rRNA, tRNA, snRNA, promoters and UTR regions. Common to all sequence types is the concept of annotating one or more feature blocks, possibly with intron blocks intervening. A feature block always uses three characters: start of block, inside block and end of block (e.g. the characters '(', 'E', and ')' for an exon block, as shown above). The advantage of using boundary markers such as '(', is that it makes it very easy to see the structure of the blocks visually and to parse it computationally. Locating areas of interest in the annotation string (and thus also in the sequence string) is as simple as performing a substring search. Alternatively, a more powerful regular expression can be used for advanced pattern matching.

THE ANNOTATION OF CHROMOSOMAL SEGMENTS

Inferring annotation from GenBank entries (1) with only a single gene or working with each gene in an entire chromosomal entry on a one-at-a-time basis is fairly simple. However, for a number of interesting studies it is useful to know about the structure and position of nearby annotated sequence regions (genes, promoters, repeats, RNAs etc.). A large part of the work that has gone into creating FeatureExtract was spent on devising a scheme for carefully annotating entire sequence

*Tel: +45 452 52489; Email: raz@cbs.dtu.dk

segments and enabling the user to co-extract annotated flanking regions relative to each main extracted sequence.

The main problem to overcome when working with flanking regions, and thus needing to annotate all features on the chromosomal segment, is that of overlapping features. Overlapping features can be artificial (e.g. a GenBank entry that uses both 'gene' and 'CDS' to annotate the same gene), but they can also represent real overlapping features (e.g overlapping genes in mitochondria or multiple putative genes defined in the same region). FeatureExtract addresses this issue by using a predefined list of features to consider for annotation in flanking regions. The list is built to minimize the problem with feature type synonyms (e.g 'CDS' versus 'gene' versus 'mRNA') but at the same time extract as much information as possible. The list can be customized.

By default, the same scheme of annotation as in the ordinary extracted sequences is used in the flanking regions, with the addition that features on the opposite strand are written in lowercase letters. For some purposes, only the presence or absence of flanking features is desirable [e.g. designing probes targeting UTR regions using OligoWiz 2.0 (2)]. Therefore, FeatureExtract has the option of a more simplistic annotation: '+' for a feature on the same strand, '−' for a feature on the opposite strand, and '#' for overlapping features. In both cases a '.' (period) is used to mark positions where no feature is present.

THE OUTPUT DATA FORMAT

A simple and very consistent scheme of a tab delimited file was chosen as the output format. The file format is intended to fulfill the following goals:

easy to parse computationally;
easy to handle in a spreadsheet or a database, or using command line tools;
consistent number of fields.

Each line in the file describes the properties of exactly one entry (Table 1) in four fields separated by tabs (Table 2):

```
Name, Sequence, Annotation, Comments.
```

'Name' is simply the name of the entry. 'Sequence' is the actual DNA sequence. 'Annotation' is the inferred annotation—guaranteed to be of the same length as the DNA sequence. 'Comments' is a free-text field; FeatureExtract concatenates all additional notes from the original GenBank data and adds additional information about the original GenBank accession ID, source (organism), type of feature extracted (e.g 'rRNA' or 'CDS'), strand ('+' or '−'), and the spliced product, if the sequence in question contains introns or frameshifts. Subfields within the comment field are separated by '/'.

As stated, an important quality of the output format is the ease with which the file can subsequently be handled computationally (e.g. by using 'grep' on the UNIX command line). A skeleton program for parsing the file could look like the following simplified example in Python to illustrate the principle:

```
for line in sys.stdin:
    tokens     = line.split("\t")
    name       = tokens[0]
    sequence   = tokens[1]
    annotation = tokens[2]
    comment    = tokens[3]
    {do computations ...}
```

The data file can easily be imported into spreadsheets such as Microsoft Excel [remember to import all fields as 'text' (3)] and databases such as Access and MySQL. Filtering the set of sequences in a spreadsheet and exporting it back in tab-delimited format is an easy way of preparing a subset of the data.

USAGE

FeatureExtract contains a diverse set of advanced options for fine-tuning the extraction. However, in most cases the default settings will be sufficient and the advanced options can safely be ignored.

The crucial step when working with the FeatureExtract web server is to specify the GenBank entries from which sequences should be extracted. The user has the option of pasting in (or uploading) a list of GenBank entry IDs or pasting in (or uploading) entire GenBank files. Hitting 'submit' at this point will run the tool with the default options and extract exon/intron annotation from all 'CDS' (protein coding genes) regions in the specified GenBank entries.

After the extraction is complete, the FeatureExtract server will provide information about the number and length of the extracted sequences and offer a link for downloading the output file.

Table 1. Example output—overall file structure

Line number	Name	Sequence	Annotation	Comment
1	alpha-D	ATGCTGACCGACTCTGACAA...	(EEEEEEEEEEEEEEEEEEE...	/gene="alpha-D"/codo...
2	alpha-A	ATGGTGCTGTCTGCCAACGA...	(EEEEEEEEEEEEEEEEEEE...	/gene="alpha-A"/codo...
3	CMGLOAD_143	ATGCTGACCGCCGAGGACAA...	(EEEEEEEEEEEEEEEEEEE...	/codon_start=1/produ...
4	CIIHBADA2_367	ATGGTGCTGTCTGCGGCTGA...	(EEEEEEEEEEEEEEEEEEE...	/note="alpha-A globi...
5	GOTHBAI_917	ATGGTGCTGTCTGCCGCCGA...	(EEEEEEEEEEEEEEEEEEE...	/note="alpha-i globi...
6	GOTHBAII_745	ATGGTGCTGTCTGCCGCCGA...	(EEEEEEEEEEEEEEEEEEE...	/note="alpha-ii glob...
7	ESGLOB01_132	ATGGTGCTGTCTGCCGCCGA...	(EEEEEEEEEEEEEEEEEEE...	/codon_start=1/produ...
8	ECPZA2GL_3481	ATGGTGCTGTCTGCCGCCGA...	(EEEEEEEEEEEEEEEEEEE...	/codon_start=1/produ...
9	AF098919_17811	ATGGCACTGACCCAAGCTGA...	(EEEEEEEEEEEEEEEEEEE...	/codon_start=1/produ...
10	AF098919_21360	ATGCTGACTGCCGAGGACAA...	(EEEEEEEEEEEEEEEEEEE...	/codon_start=1/produ...
11	AF098919_24360	ATGGTGCTGTCCGCTGCTGA...	(EEEEEEEEEEEEEEEEEEE...	/codon_start=1/produ...

Each line contains four tab-separated fields (Name, Sequence, Annotation and Comments) representing an individual feature. In this example the features extracted are protein coding genes (CDS) from the following GenBank entries: AB001981, X01831, J00923, J00043, J00044, X01086, X07053, AF098919. For readability the fields have been truncated after 20 letters.

Table 2. Example output—field details

Field	Contents
Name	alpha-D
Sequence	ATGCTGACCGACTCTGACAAGAAGCTGGTCCTGCAGGTGTGGGAGAAGGTGATCCGCCACCCAGACTGTGGAGCCGAGG CCCTGGAGAGGTGCGGGCTGAGCTTGGGGAAACCATGGGCAAGGGGGGCGACTGGGTGGGAGCCCTACAGGGCTGCTGG GGGTTGTTCGGCTGGGGGTCAGCACTGACCATCCCGCTCCCGCAGCTGTTCACCACCTACCCCCAGACCAAGACCTACT TCCCCCACTTCGACTTGCACCATGGCTCCGACCAGGTCCGCAACCACGGCAAGAAGGTGTTGGCCGCCTTGGGCAACGC TGTCAAGAGCCTGGGCAACCTCAGCCAAGCCCTGTCTGACCTCAGCGACCTGCATGCCTACAACCTGCGTGTCGACCCT GTCAACTTCAAGGCAGGCGGGGGACGGGGGTCAGGGGCCGGGGAGTTGGGGGCCAGGGACCTGGTTGGGGATCCGGGGC CATGCCGGCGGTACTGAGCCCTGTTTTGCCTTGCAGCTGCTGGCGCAGTGCTTCCACGTGGTGCTGGCCACACACCTGG GCAACGACTACACCCCGGAGGCACATGCTGCCTTCGACAAGTTCCTGTCGGCTGTGTGCACCGTGCTGGCCGAGAAGTA CAGATAA
Annotation	(EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE EEEEEEEEE)DIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIA(EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE EEEEEEEEEEE)DIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIA(EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE EEEEEE)
Comment	/gene="alpha-D"/codon_start=1/product="alpha-D globin"/protein_id="BAA19668.1"/ db_xref="GI:1943997"/translation="MLTDSDKKLVLQVWEKVIRHPDCGAEALERLFTTYPQTKTYFPHF DLHHGSDQVRNHGKKVLAALGNAVKSLGNLSQALSDLSDLHAYNLRVDPVNFKLLAQCFHVVLATHLGNDYTPEAHAAF DKFLSAVCTVLAEKYR" /Genbank_acc="AB001981"; /Source="Columba livia (domestic pigeon)"; /feature_type="CDS"; /strand="+"; /spliced_product="atgctgaccgactctgacaa gaagctggtcctgcaggtgtgggagaaggtgatccgccacccagactgtggagccgaggccctggagaggctgttcacc acctacccccagaccaagacctacttcccccacttcgacttgcaccatggctccgaccaggtccgcaaccacggcaaga aggtgttggccgccttgggcaacgctgtcaagagcctgggcaacctcagccaagccctgtctgacctcagcgacctgca tgcctacaacctgcgtgtcgaccctgtcaacttcaagctgctggcgcagtgcttccacgtggtgctggccacacacctg ggcaacgactacaccccggaggcacatgctgccttcgacaagttcctgtcggctgtgtgcaccgtgctggccgagaagt acagataa"; /spliced_annotation="(EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE)(EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE EEEEEEEEE)(EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE)";

Detailed example of data extracted from the GenBank entry AB001981 (first CDS).

Full documentation of options (basic and advanced), the output file format, examples of usage and sample input data are to be found at the FeatureExtract website.

ACKNOWLEDGEMENTS

FeatureExtract is inspired by programs and concepts developed by Søren Brunak, Kristoffer Rapacki and Lars Juhl Jensen. The author would like to thank Anders Gorm Pedersen and Thomas Skøt Jensen for comments on the manuscript. A grant from the Danish Technical Research Council (STVF) for 'Systemic Transcriptomics in Biotechnology' financed this work. Funding to pay the Open Access publication charges for this article was provided by a grant from the Danish Technical Research Council (STFV) for 'Systemic Transcriptomics in Biotechnology'.

Conflict of interest statement. None declared.

REFERENCES

1. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2004) GenBank: update. *Nucleic Acids Res*., **32**, D23–D26.
2. Wernersson,R. and Nielsen,H.B. (2005) OligoWiz 2.0—integrating sequence feature annotation into design of microarray probes. *Nucleic Acids Res*., **33**, W611–W615.
3. Zeeberg,B.R., Riss,J., Kane,D.W., Bussey,K.J., Uchio,E., Linehan,W.M., Barrett,J.C. and Weinstein,J.N. (2004) Mistaken Identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics. *BMC Bioinformatics*, **5**, 80.